Statistics Canada Statistique Canada

Methodology Branch

Census & Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes de recensement
et d'enquêtes ménages

Canada

Z-151F
c.2

METHODES D'ESTIMATION REGIONALE
À STATISTIQUE CANADA

Communication présentée par Statistique Canada
à la conférence des statisticiens européens
Réunions sur la méthodologie statistique, février 1986

Ce document a été préparé par J. Douglas Drew
qui a utilisé les travaux d'un grand nombre de personnes à Statistique Canada
CHSM 85-078 F

MÉTHODES D'ESTIMATION RÉGIONALE À STATISTIQUE CANADA

1. Au cours des dernières années, on a observé une augmentation de la demande de données régionales au Canada. Les administrations se sont préocupées davantage des problèmes de distribution, de justice et de disparité, et des programmes ont été préparés pour aider les économiquement faibles, qu'il s'agisse de régions géographiques ou de sous-groupes de la population, et ces programmes ont été par la suite suivis. De plus, comme beaucoup de décisions d'entreprises sont prises en fonction des conditions sociales, économiques et environnementales au niveau local, la demande de statistiques régionales de la part du secteur privé est élevée. Enfin, les politiciens désirent disposer de statistiques sur la situation relative de leurs électeurs par rapport aux autres.

2. En même temps que cette demande augmentait, les progrès technologiques et statistiques enregistrés ont conduit à l'élaboration de données régionales. Ces progrès comprennent l'automatisation des systèmes de données administratives, l'utilisation accrue du codage postal comme identificateur géographique de ces fichiers, la réduction du coût des calculs en raison de l'existence d'ordinateurs plus puissants et l'élaboration continue de méthodes de modélisation statistique qui permettent le calcul d'estimations régionales à partir simultanément de données d'enquête et de recensement ou de dossiers administratifs, qui seront produits à un niveau inférieur à celui définissant la limite de fiabilité d'un échantillon d'enquête.

3. En 1982, Statistique Canada a commencé à mettre au point un projet d'élaboration et de diffusion systématique et intégré de données régionales, dans le cadre duquel ces dernières étaient considérées dans une perspective géographique, la région étant l'unité à décrire et les différents paramètres en constituant le détail.

4. Le programme des données régionales a été lancé en 1983, avec une période d'élaboration de trois années, après laquelle le programme deviendra dans une large mesure autonome grâce aux recettes provenant de la vente de produits sur les données régionales. En plus de ce travail d'élaboration et du rôle de coordination à Statistique Canada, le programme était également considéré comme un moyen de coordonner les autres centres d'intérêt des programmes statistiques de données régionales existant dans d'autres ministères fédéraux et les provinces. Une description des questions de politique et une vue générale du programme figurent dans Brackstone (1985).

5. On a identifié trois composantes du programme: l'élaboration des données, les systèmes de données et l'infrastructure. Les sections ultérieures de cette communication portent sur quatre éléments de l'élaboration des données: les données du marché du travail, les données des entreprises, les données du revenu des familles et les estimations démographiques postcensitaires. Le reste de cette section contient une brève description des autres éléments du programme.

a) Systèmes de données

6. Un élément fondamental du système de données régionales est un répertoire des ensembles de données régionales existantes, qui comprend à la fois leurs caractéristiques et les moyens d'y accéder. Les ensembles de données elles-mêmes peuvent être classés soit comme des bases de données sources ou comme des bases de données sommaires. Les bases sources contiennent habituellement des micro-données, qui se trouvent normalement chez les unités de collecte des données. L'accès y est habituellement limité pour des raisons de confidentialité. Par contre, les bases de données sommaires constituent les consolidations de données provenant d'une foule de sources à un niveau géographique unique. À partir des bases de données sommaires, il est possible de tirer des séries de produits qui peuvent être préparés pour répondre aux besoins d'un utilisateur, comme par exemple des relevés de données sur des caractéristiques particulières, les données sur disquettes, etc.

b) Infrastructure

7. L'élément le plus important de l'infrastructure est la géographie. Le programme des données régionales s'efforce de fournir des données dépassant le cadre des régions classiques. On peut définir des régions non-standard en termes de critères géographiques normalisés pour lesquels il est possible d'obtenir des données grâce au géocodage ou par l'emploi de codes postaux. Le géocodage consiste à identifier les coordonnées de longitude et de latitude et il est utilisé depuis le recensement de la population de 1971. La méthode du code postal utilise un fichier de conversion pour convertir les codes postaux en codes géographiques normalisés. Les deux méthodes sont précises au niveau des blocs de rues dans les régions urbaines, mais moins dans les régions rurales. Un élément important du programme des données régionales a été et demeurera l'amélioration de ces méthodes. C'est ainsi que lors du recensement de 1986, le code postal sera saisi pour un échantillon de 20% des logements recensés, afin de perfectionner le processus de conversion.

II. ÉLABORATION DES DONNÉES DU MARCHÉ DU TRAVAIL

a) Objectifs

8. Les objectifs de ce programme étaient d'examiner d'éventuels prolongements de l'ensemble des mesures du marché du travail afin de les étendre aux petites régions et de préparer des recommandations sur les données à produire par Statistique Canada. Les premiers travaux ont porté principalement sur les estimations des divisions de recensement, qui sont des régions géographiques normalisées se situant immédiatement au-dessous du niveau auquel les données sont actuellement publiées.

b) Sources

9. Cette section traite des mesures du marché du travail basées selon le lieu de résidence (obtenues à partir des ménages ou des personnes), tandis que les mesures basées sur le lieu de travail (recueillies auprès des entreprises) sont examinées dans la section suivante.

10. Les principales sources de données régionales sur le marché du travail sont l'enquête sur la population active, les recensements de la population et les données administratives. Les avantages et les inconvénients de chacune sont examinés ci-après.

11. L'enquête sur la population active est une enquête mensuelle auprès de 51,000 ménages touchant la population civile hors institutions des dix provinces du Canada. Un nouvel échantillon a été mis en oeuvre en janvier 1985, qui a entraîné une augmentation de la quantité et de la qualité des données infraprovinciales obtenues de l'enquête. Avant, l'enquête était conçue principalement dans le but de fournir les données nationales et provinciales les plus fiables. Comme l'objectif principal de la révision était l'améliortion des données infraprovinciales, des données sont maintenant publiées chaque mois pour 66 régions infraprovinciales et 24 régions métropolitaines de recensement (RMR) avec un CV pour les chômeurs de 15%, comparativement à 53 régions et 23 RMR auparavant. Une lacune comportante de l'EPA pour ce qui est de ses possibilités régionales est sa capacité réduite de produire des estimations mensuelles pour des régions plus petites en raison de la petite taille des échantillons en cause. Il est possible de régler ce problème dans un certaine mesure en accumulant des échantillons mensuels pour obtenir des données moyennes trimestrielles, annuelles ou pluriannuelles. Ainsi, par exemple, avec le nouvel échantillon, des données trimestrielles sont maintenant produites pour 42 villes hors RMR, avec un CV moyen de 20% pour les chômeurs.

12. Au Canada, le recensement de la population et des logements est effectué tous les cinq ans. Lors des derniers recensements, les reseignements sur l'emploi et le chômage ont été recueillis grâce à un échantillon de 20%, comme ce sera le cas également lors du recensement de 1986. Les principaux avantages des données du recensement dans une perspective régionale sont leur souplesse de totalisation non seulement pour des régions infraprovinciales normalisées telles que les divisions de recensement et les circonscriptions électorales fédérales, mais également pour des régions précisées par l'utilisateur, et des erreurs d'échantillonnage relativement petites (ainsi, un CV moyen de 5% pour les estimations des chômeurs au niveau de la division du recensement). Leurs inconvénients résident dans leur fréquence réduite et le délai relativement long pour la publication des données (environ un an).

13. Les renseignements administratifs les plus appropriés pour l'estimation du chômage proviennent des dossiers de l'assurance-chômage (AC). Le concept du chômage de l'enquête sur la population active fait la distinction entre deux composantes: les personnes qui perdent/quittent un emploi, et les nouveaux entrants ou les rentrants dans la population active. Le nombre de bénéficiaires de l'AC normaux sans gain correspond par son concept d'assez près à la composante des personnes ayant perdu ou quitté leur emploi, qui en règle générale interviennent pour 70% du nombre des chômeurs. Statistique Canada produit des chiffres directs des prestataires normaux sans gain au niveau de la division de recensement. De plus, Statistique Canada produit le ratio AC/P qui est le ratio des bénéficiaires normaux sans gain à la population en âge de travailler (15-64 ans). Les avantages des données administratives sont leur fréquence et les délais. Elles se caractérisent également par une souplesse géographique considérable, bien que limitées dans le cas des petites régions par l'erreur provenant de la conversion géographique des codes postaux, qui peut poser des problèmes. Les inconvénients comprennent l'absence de cohérence avec les concepts, le champ d'observation de l'enquête et le risque d'interruption de la continuité des séries chronologiques en raison de changement dans le programme administratif.

c) Estimateurs régionaux

14. L'objectif des travaux d'élaboration des données sur le marché du travail était d'étendre le champ des données régionales, tout en garantissant i) la cohérence avec les concepts de l'EPA et ii) l'agrégation aux estimations de l'EPA déjà publiées à des niveaux géographiques plus élevés. Les techniques d'estimations régionales considérées peuvent en gros être classées soit comme des estimateurs basés sur le plan de sondage, des estimateurs basés sur un modèle ou des estimateurs combinés. Ces estimateurs sont examinés ci-dessous, avec un brève mention de quelques travaux antérieurs au programmes des données régionales.

15. Les estimateurs basés sur le plan de sondage comprennent les estimateurs directs (ou de domaine ou à expansion simple) et les estimateurs de domaine post-stratifés. L'estimateur direct s'obtient en utilisant la procédure d'estimation normale convenant au plan de l'enquête aux unités d'échantillonnage faisant partie de chaque région ou domaine. Un problème avec l'estimateur direct est l'absence générale de coincidence des limites des strates d'enquête et des limites des régions étudiées, problème qui se trouve aggravé par la nature en grappes des plans de sondage utilisés dans la pratique. Cette situation peut se traduire par un sur ou un sous-échantillonnage de la petite région, de sorte que les estimations des totaux pour un échantillon donné peuvent être gravement biaisées. Lorsqu'il est possible d'obtenir de sources extérieures des renseignements étalons sur une variable connexe (population de la région, par exemple), ce problème peut être réglé par la post-stratification.

16. Les estimateurs basés sur un modèle comprennent les estimateurs synthétiques et les estimateurs à maintien de la structure (EMS) de Purcell et Kish (1979). Les estimateurs synthétiques ont d'abord été étudiés pour la première fois à Statistique Canada par Ghangurde et Singh (1972) dans le contexte de calcul d'estimations régionales du marché du travail. Leur estimateur répartissait les estimations d'enquête (groupe d'âge/sexe) pour les régions plus grandes aux petites régions à partir de la part que possédait chaque petite région de la population de la région plus grande (par groupe d'âge/sexe) lors d'un recensement récent. Bien que les estimateurs synthétiques aient donné une réduction appréciable de l'erreur d'échantillonnage, il y a une hypothèse implicite d'homogénéité au sein du sous-groupe de la population entre la petite région et la grande région, qui, si elle ne se vérifie pas, peut entraîner un biais sérieux.

17. Les estimateurs combinés essaient de combiner les deux autres types d'estimateurs de façon à en bénéficier des avantages et à en éliminer les inconvénients. Drew, Singh et Choudhry (1982) ont proposé et évalué un estimateur qu'ils ont baptisé "estimateur dépendant de l'échantillon", qui repose uniquement sur l'estimateur post-stratifié lorsque l'échantillon de la petite région est égal ou supérieur à celui prévu par le plan de sondage, mais qui autrement passe à un combination linéaire d'estimateurs post-stratifiés et synthétiques, avec une augmentation des poids de la composante synthétique, à mesure que la petite région devient de plus en plus sous-échantillonnée. Dans une étude de Monte Carlo, on a constaté que l'estimateur dépendant de l'échantillon avait une e.q.m. comparable, mais un biais plus petit, que l'estimateur synthétique. On avait recommandé alors d'utiliser l'estimateur dépendant de l'échantillon en combinaison avec un mise en moyenne des données mensuelles afin de produire des estimations moyennes annuelles ou pluriannuelles fiables pour les divisions de recensement.

d) Élaboration des données sur le marché du travail en vertu du programme des données régionales

18. En vertu du programme des données régionales, on a identifié deux directions pour l'élaboration des données. La première consistait en la mise en oeuvre d'une estimation dépendant de l'échantillon, se rattachant au départ à la production d'estimations de l'emploi et du chômage pour les divisions de recensement, mais avec la possibilité de généraliser plus tard de façon à pouvoir absorber des systèmes régionaux souples. La deuxième direction consistait à poursuivre les recherches dans le domaine des estimateurs de modèles, qui pourraient dans un avenir plus ou moins rapproché être améliorés et remplacer ainsi l'estimateur dépendant de l'échantillon.

19. Cette stratégie s'est révélée nécessaire afin de répondre aux besoins immédiats des utilisateurs, tout en permettant la poursuite des travaux de recherche avec d'intéressantes perspectives. La demande la plus importante provenait du ministère de l'Expansion régionale et industrielle, qui demandait des estimations moyennes sur trois ans des taux de chômage et des ratios emploi/population pour des divisions de recensement individuelles ou combinées. Ces estimations sont maintenant produites couramment et servent à la préparation d'un indice de développement annuel, dans lequel les divisions de recensement sont classées en quatre niveaux se qualifiant pour des niveaux d'assistance maximum de plus en plus élevés des programmes de développement industriel approuvés.

20. En ce qui concerne d'autre recherches, Choudry et Bélanger (1985) ont entrepris une étude de Monte Carlo dans laquelle le plan de sondage de l'EPA était simulé pour une province du Canada en utilisant les données du recensement. Des estimateurs directs, post-stratifiés, dépendant de l'échantillon, synthétiques et EMS ont été évalués. Leur formulation de l'estimateur synthétique était la même que celle de Ghangurde et Singh (1972), tandis que l'estimateur EMS commençait par des estimations EPA par groupes d'âge/sexe pour les petites régions provenant du recensement précédant comme structure d'association; on a procédé à une estimation itérative par le quotient utilisant les totaux marginaux suivants: estimations d'enquête selon l'âge/sexe pour la région économique, et population totale pour la division du recensement.

21. Ils devaient conclure que les méthodes synthétiques et EMS ont des biais relatifs assez importants pour les chômeurs (7.7% et 11.3% respectivement), mais que par contre leur erreur d'échantillonnage (coefficient de variation) était inférieure de 50% à celle de l'estimateur dépendant de l'échantillon.

22. Il découle de ces conclusions que l'estimateur dépendant de l'échantillon a une e.q.m. inférieure à celle des estimateurs basés sur un modèle pour les estimations des divisions de recensement lorsque l'on fait la moyenne sur une période de trois ans, comme c'est le cas de l'indice de développement. Ceci s'explique par le fait que les erreurs d'échantillonnage se trouvent réduites lorsqu'on fait la moyenne sur une longue période, au point où le biais élevé des estimateurs basés sur un modèle domine. Pour les périodes inférieures à un an cependant, les variances dominent les e.q.m. et les estimateurs basés sur le modèle donnent par conséquent de meilleurs résultats.

23. Comme le biais joue un rôle essentiel dans le choix entre différents estimateurs régionaux, une des directions qui s'ouvre à la recherche sera une étude en profondeur de celui-ci. La base géographique de l'étude sera étendue de façon à regrouper toutes les provinces, et les effets de différents renseignements auxiliaires seront étudiés. Parmi ceux-ci figurent l'utilisation des données de l'AC, l'utilisation des chiffres du recensement selon l'âge/sexe par type de région (rurale/urbaine) et l'utilisation des données de l'AC pour l'estimation de la composante "personnes ayant perdu/quitté leur emploi" chez les chômeurs et les données du recensement pour l'estimation de la composante entrants/rentrants. Les estimateurs qui donnent les meilleurs résultats en termes de biais peuvent ensuite être inclus dans une autre étude de Monte Carlo pour évaluer l'e.q.m.

24. Une autre direction possible pour les recherches serait l'étude des estimateurs de régression. Les premières études des modèles de régression logistique pour estimer le chômage au niveau de la région en utilisant les chiffres de l'AC et des estimations de population comme variables indépendantes n'ont pas donné de résultats prometteurs. On envisage d'étudier un estimateur de régression plus perfectionné qui incorporerait comme variables indépendantes le ratio logarithmique des bénéficiaires de l'AC à la population, la branche d'activité et le degré d'urbanisation selon le recensement ainsi que des variables dichotomiques pour le mois et l'année, avec des ajustements normalisés pour prendre en compte l'auto-corrélation des erreurs d'échantillonnage d'un mois à l'autre.

III. ÉLABORATION DES DONNÉES RÉGIONALES SUR LES ENTREPRISES

25. Suite aux travaux d'élaboration des statistiques sociales, un projet a été entrepris pour élaborer des données régionales sur les entreprises. Les objectifs précis étaient les données pour les divisions de recensement, avec une ventilation selon le groupe industriel (18) et la taille de l'établissement (2-4 catégories), pour les variables suivantes: revenu brut des entrepriss (RBE), salaires et traitements (ST), emploi et nombre d'établissements.

26. Le cadre général de l'élaboration de données régionales sur les entreprises n'est pas sans rappeler celui des données régionales du marché du travail décrites à la section précédente. Des données d'enquêtes existent pour un échantillon de la population, avec des renseignements détaillés basés sur les concepts établis, et en règle générale les régions intéressantes sont les domaines c'est-à-dire qu'elles ne sont pas prises en compte dans les strates du plan du sondage. Il existe des données administratives qui peuvent être géocodées pour les régions grâce aux codes ou aux adresses postales, et leur champ d'observation et leurs concepts diffèrent dans une certaine mesure de ceux qui seraient souhaitables.

27. Compte tenue de la similitude du cadre, il n'est donc pas surprenant qu'un grand nombre des questions à régler soient semblables à celles examinées à la section précédente. La difficulté principale supplémentaire est que dans le cas de la statistique des entreprises, l'unité qui fournit le renseignement peut différer sensiblement de celle pour laquelle on désire des statistiques. Ainsi dans le cas des entreprises constituées, des données administratives peuvent être fournies au niveau de la société, alors que les données régionales conviennent mieux lorsqu'elles se rapportent au lieu de l'activité marchande.

28. Un résumé des travaux d'élaboration suit, présenté de façon plus détaillée par Sande (1985). L'estimation de RBE et de ST est examinée d'abord pour les entreprises individuelles, et ensuite pour les entreprises constituées, ce qui comprend un examen de la désagrégation des sociétés en unités déclarantes d'emploi (UDE). Enfin, l'estimation de l'emploi est examinée.

a) Données financières sur les entreprises individuelles

29. Les dossiers de Revenu Canada, auxquels Statistique Canada a accès en vertu de la Loi sur la statistique de 1971, sont la source des renseignements financiers sur les entreprises individuelles et constituées. Dans le premier cas, chaque personne qui a une entreprise doit présenter un état financier avec sa déclaration fiscale. Un échantillon de ces dernières est envoyé à Statistique Canada où les données financières sont transcrites et un code de branche d'activité est ajouté, pour servir ensuite à une foule d'enquêtes. Le plan de sondage est simple, la seule variable de stratification étant le RBE. Il y a un échantillon à 10% des déclarants dont le RBE est compris entre $10,000 et $25,000, un échantillon à 25% pour ceux dont le RBE est compris entre $25,000 et $500,000, et un échantillon à 100% pour les entreprises avec un RBE dépassant $500,000. En 1982, la taille de l'échantillon était de 150,000 sur une population totale de 500,000.

30. En plus des données d'échantillon, une autre source de renseignements existe dans le cas des déclarants dont le RBE est compris entre $25,000 et $500,000, sous la forme d'un fichier maintenu par Revenu Canada aux fins de vérification. En 1981, le fichier de vérification contenait des renseignements sur le total des ventes qui étaient assez semblables par leur concept au RBE, mais le fichier ne contenait rien d'équivalent à ST. Les problèmes qui se posent pour l'utilisation de ce fichier sont les changements fréquents de sa structure au cours des dernières années. Il contient maintenant moins de renseignements qu'auparavant.

31. L'élaboration des données a procédé, au départ, au calcul d'estimations directes de RBE et de ST à partir des données de l'échantillon. Deux problèmes importants se sont posés:

i) Le géocodage automatisé des fichiers, reliant les adresses et/ou les codes postaux des entreprises aux divisions de recensement par des fichiers de conversion des adresses aux codes postaux et des codes postaux aux divisions de recensement. Cette opération a été très réussie, et un très petit nombre de cas ont nécessité une intervention manuelle.

ii) Les données manquantes. Dans le passé, lors de la transcription des données de Revenu Canada sur les salaires et les traitements, on ne faisait aucune distinction entre les entrées nulles et les non-réponses. On étudie actuellement la façon de régler ce problème à l'étape de l'imputation.

32. Des études ont également été effectuées afin de trouver des moyens d'utiliser le fichier de vérification avec certaines hypothèses simplificatrices pour améliorer l'estimation basée sur l'échantillon pour les entreprises dont le RBE est compris entre $25,000 et $500,000 (Hidiroglou 1985). Les étapes de ces études consistent en: i) la comparabilité des données du fichier de vérification et de celles du fichier échantillon, et ii) les résultats d'autres estimateurs régionaux. Les principaux résultats sont examinés ci-dessous.

33. Voici les résultats de la comparaison des données.

i) Données géographiques. L'adresse du déclarant dans le fichier de vérification devait se révéler être une bonne approximation de l'adresse de l'entreprise (qui n'existe que dans le fichier échantillon), avec une concordance à 95% lorsqu'on utilisait les différentes adresses pour attribuer les divisions de recensement.

ii) Données industrielles. Il existe une concordance relativement bonne (78%) entre les codes affectés par Statistique Canada pour l'échantillon et par Revenu Canada pour le fichier de vérification pour les dix-huit grandes divisions. On a suggéré des méthodes pour réduire les différences.

iii) Données financières. On avait constaté que les données des ventes du fichier de vérification concordaient de près avec celles du fichier échantillon, les divergences se situant dans un intervalle de 10% pour plus de 90% des cas.

34. Dans la deuxième phase, on a envisagé l'estimation des ST présents dans le fichier d'échantillonnage seulement, en utilisant le fichier de vérification comme source de renseignements auxiliaires sur le nombre d'entreprises et le RBE. Des variables présentes dans les deux fichiers, on devait constater que RBE était le meilleur prédicteur de ST. Une étude de Monte Carlo des estimateurs de sondage, basés sur un modèle et combinés devait révéler que les estimateurs combinés étaient les meilleurs en termes de biais et d'erreur quadratique moyenne minimums, que les estimateurs basés sur le modèle (synthétiques) pouvaient être très biaisés, que l'estimateur direct avait l'erreur quadratique moyenne la plus élevée et que la performance des estimateurs post-stratifiés n'était que très légèrement inférieure à celle des estimateurs combinés.

b) Données financières sur les entreprises individuelles

35. Les principales sources de données sont:

i) Les données de Revenu Canada: Il existe un échantillon détaillé de 25,000 dossiers, à partir duquel ST et RBE sont imputés grâce à un modèle à ratio faisant intervenir l'actif, les ventes et d'autres renseignements pour l'univers des 500,000 entreprises.

ii) Le fichier supplémentaire T4: Le feuillet T4 au Canada est le résumé annuel de la rémunération et des déductions que chaque contribuable reçoit de son employeur. Chaque dossier T4 contient un numéro de déduction de paye (DP) qui peut servir à identifier l'employeur.

iii) Le registre des entreprises (RE): Statistique Canada tient un registre des entreprises, qui est construit suivant le système DP. Il raccorde les numéros DP aux établissement et aux sociétés, et il est également raccordé aux bases de sondage des principales enquêtes économiques. La géographie et la branche d'activité y sont codées.

36. La difficulté principale dans le cas des données sur les sociétés de Revenu Canada est qu'elles portent sur la siège social et non sur le lieu proprement dit de l'actvité économique. D'autres moyens de désagréger les données sur les sociétés sont à l'étude. Une approche consisterait à utiliser les données T4 pour obtenir ST par numéro DP et ensuite raccorder le numéro DP aux établissements ou aux UDE grâce au RE et à l'enquête sur l'emploi, la rémunération et les heures de travail. Les numéros DP raccordés à plus d'une UDE sur ERHT verront une affectation proportionnelle de ST à l'emploi déclaré. Une autre approche dans le cas de ST est le raccordement du fichier imputé de Revenu Canada directement au RE. Dans le cas de la désagrégation de RBE, l'utilisation de recensements économiques est à l'étude, puisque cette façon d'agir permettrait d'utiliser les données RBE du recensement comme base pour l'affectation du total des recettes provenant du fichier T2 entre les établissements constituants.

c) Estimations de l'emploi

37. L'enquête sur l'emploi, la rémumération et les heures de travail est une enquête mensuelle qui recueille des données de 70,000 unités déclarantes d'emploi. L'échantillon est stratifié par province, taille d'entreprise et grande division industrielle.

38. Les premiers efforts portent sur la production d'estimations directes et post-stratifiées de l'emploi par division industrielle pour les divisions de recensement et sur des estimations de leurs variances. Des méthodes pour regrouper les branches d'activité et les DR sont également à l'étude. Compte tenu des résultats, on pourrait envisager plus tard des estimateurs de modèle ou combinés plus perfectionnés comme ceux qui ont été examinés dans les sections précédentes, et la production de moyennes annuelles.

IV. STATISTIQUES SUR LES FAMILLES ET LE REVENU DES FAMILLES

39. Les sources actuelles des statistiques infraprovinciales du revenu des familles sont le recensement de la population et l'enquête annuelle sur les finances des consommateurs (EFC). Les données sur le revenu sont recueillies par sondage pour chaque recensement décennal et aussi pour quelques recensements quinquennaux, ce qui sera le cas en 1986. En 1981, avec un échantillon à 20%, le CV moyen du revenu annuel des divisions de recensement était de 1.5%. A partir de l'EFC, on ne produit actuellement pas un grand nombre de données infraprovinciales. Des estimations sont publiées pour certaines RMR. Pour les seize RMR publiées en 1982, le CV du revenu moyen de la famille était de 3.6%.

40. Les travaux de recherches ont porté principalement sur l'examen des possibilités du fichier du revenu des particuliers de Revenu Canada pour des statistiques sur les familles et le revenu familial. Les fichiers de Revenu Canada sont des fichiers de personnes. Ils contiennent le numéro d'assurance sociale (NAS), l'état matrimonial, le NAS du conjoint, les sources de revenu et le montant des déductions et exemptions fiscales.

41. Lors de l'identification des familles à partir des fichers de Revenu Canada, on a essayé de suivre d'aussi près que possible la définition du recensement de la famille, qui correspond à un époux et une épouse ou à un parent seul, avec tout enfant non marié vivant dans la même ménage. Les conjoints qui ont rempli tous les deux des déclarations d'impôt ont été raccordés en utilisant le NAS du conjoint lorsqu'il y en avait un, ou à partir du nom, de l'adresse, de l'âge et de l'état matrimonial. On a également essayé de raccorder les contribuables célibataires de moins de 30 ans à une famille à partir du nom et de l'adresse. Enfin, pour chaque famille, le nombre d'enfants non déclarants a été estimé à partir des montants des différents avantages fiscaux réclamés.

42. Des estimations expérimentales des familles ont été produites pour la plus petite province canadienne et évaluées à partir des données du recensement (Auger 1985). Ces comparaisons ont révélé que les estimations administratives du nombre de familles et de personnes seules n'étaient pas trop différentes de celles du recensement, avec une sous-estimation de 6% et une surestimation de 9% respectivement. L'énumération des enfants de moins de 18 ans devait se révéler proche de 100%, tandis que celui pour les enfants de 18 ans et plus était plus proche de 80%. Ce dernier résultat s'explique probablement par les limites des méthodes informatiques utilisées dans l'appariement des enfants qui ont présenté une déclaration fiscale. Toutefois, ces procédures ont été améliorées. L'un des principaux problèmes lié à ce fichier administratif devait se révéler être la sous-énumération des personnes âgées de 65 ans et plus. De même, le nombre estimatif de personnes seules de moins de 35 ans devait se révéler beaucoup plus élevé que dans le cas du recensement, ce qui s'explique principalement par les couples en union libre, qui n'ont pas été considérés comme des familles époux-épouse dans l'étude, et à la présence d'enfants non appariés.

43. A partir des possibilités prouvées d'estimer les familles et le revenu des familles (principalement pour celles dont les membres sont âgés de moins de 65 ans), des travaux sont actuellement en cours pour obtenir des estimations des familles pour toutes les provinces, comprenant un appariement des unions libres. Enfin, quelques travaux distincts ont été entrepris avec le fichier des pensions de sécurité de la vieillesse, ce qui servira à améliorer l'énumération des personnes âgées et de leurs familles.

V. ESTIMATION DÉMOGRAPHIQUE POSTCENSITAIRE

44. Avant 1982, Statistique Canada préparait des estimations démographiques annuelles pour les divisions de recensement et les RMR en utilisant une méthode d'estimation par les composantes. Ces données étaient publiées de 15 à 18 mois après la période de référence.

45. Des recherches ont été effectuées en vertu du programme des données régionales pour évaluer d'autres procédures d'estimation afin d'améliorer la précision et de réduire les délais. Ces travaux se sont traduits par l'élaboration et la mise en oeuvre d'une procédure en vertu de laquelle deux ensembles d'estimations seront publiés chaque année pour les divisions de recensement et les RMR, 3-4 mois et 11-15 mois respectivement après la période de référence. Verma et Basavarajappa (1985) ont présenté les résultats de leurs recherches. Voici quelques faits saillants.

46. En vertu de la nouvelle approche, les estimations de l'année courante sont basées sur les estimations des composantes de l'année précédente, auxquelles on ajoute une estimation de la variation de la population, basée sur la différence entre les estimations de régression pour les deux dates considérées. Les estimations ainsi obtenues ont été désignées "estimations emboîtées". Pour la même période de référence, des estimations sont produites plus tard en utilisant une approche par composantes qui donne des renseignements sur les composantes des variations, c'est-à-dire les naissances, les décès et la migration nette.

47. Les variables symptomatiques utilisées dans l'estimation de régression différaient d'une province à l'autre suivant l'existence de dossiers administratifs différents, et comprenaient des données sur les allocations familiales, des chiffres provenant des fichiers provinciaux sur les soins de santé et les branchements électriques. La méthode des composantes utilise les naissances et les décès des registres d'état civil et la migration estimée à partir des fichiers fiscaux des particuliers de Revenu Canada.

48. Les études d'évaluation ont permis de constater que les estimations des régressions emboîtées sont meilleures que celles de la méthode des composantes ou de la régression normale pour l'estimation de la population totale, avec une erreur absolue moyenne de 1.7% pour les divisions de recensement. La stratégie d'estimation adoptée donne donc des estimations courantes et précises du total de la population, des estimations des composantes des variations de la population devenant disponibles plus tard avec les estimations des composantes.

49. D'autres travaux de recherches sont prévus pour évaluer l'effet d'une transformation log-linéaire du modèle et de tester de façon plus poussée le modèle avec différents ensembles de fichiers de données symptomatiques, tels que ceux des soins de santé et des permis de conduire.

VI. CONCLUSIONS

50. A mesure que sa troisième année d'existence tire à sa fin, le programme des données régionales a atteint la plus grande partie de ses objectifs initiaux d'élaboration. En dépit d'un accent plus grand lors de la deuxième phase du programme sur la production et le recouvrement des coûts, il existe encore d'importantes possibilités d'élaboration, et une partie du programme continuera d'y être consacrée.

51. Ces possibilités comprennent l'exploitation plus grande de données provenant d'enquêtes autres que l'enquête sur la population active et l'enquête sur l'emploi, la rémunération et les heures de travail, d'autres recherches en techniques afin de garantir la confidentialité, en particulier pour les données régionales des entreprises, et l'acquisition de nouvelles sources de données administratives telles que les dossiers de la sécurité de vieillesse, les fichiers des branchements téléphoniques et électriques, les immatriculations de voitures automobiles, et les rôles d'impôt des municipalités.

BIBLIOGRAPHIE

Auger, E. (1985): Statistiques sur les familles et leur revenu à partir du fichier d'impôt personnel de Revenu Canada: une étude-pilote. Rapport présenté au symposium international sur la statistique des petites régions, Ottawa, mai 1985.

Brackstone, G.J. (985): Small Area Data: Policy Issues and Technical Challenges. Rapport présenté au symposium international sur la statistique des petites régions, Ottawa, Canada, mai 1985.

Choudhry, G.H. and Bélanger, Y. (1985): Small Area Estimates from Sample Surveys. Rapport présenté au symposium international sur la statistique des petites régions, Ottawa, Canada, mai 1985.

Drew, J.D., Singh, M.P. et Choudhry, G.H. (1982): Evaluation des techniques d'estimation pour les petites régions dans l'enquête sur la population active au Canada, Techniques d'enquête, 8, 19-52.

Ghangurde, P.D. and Singh, M.P. (1973): Synthetic Estimation in Periodic Household Surveys, Techniques d'enquête, 3, 152-181.

Hidiroglou, M.A. (1985): Problems Associated with the Estimation of Small Area Business Data. Rapport présenté à la réunion de la Société Statistique du Canada, Winnipeg, Canada, juin 1985.

Purcell, N.J., et Kish, L. (1979): Estimation for Small Domains, Biometrics, 35, 365-384.

Sande, I. (1985): The Small Area Business Data Development Project at Statistics Canada. Rapport présenté au symposium international sur la statistique des petites régions, Ottawa, Canada, mai 1985.

Verma, R.B.P. et Basavarajappa, K.G. (1985): Recent Developments in the Estimation of Population for Small Areas in Canada by Regression Method. Rapport présenté au symposium international sur la statistique des petites régions, Ottawa, Canada, mai 1985.